JoVA: Unified Multimodal Learning for Joint Video-Audio Generation and Editing

Xiaohu Huang[1*], Haoyang He[2*], Hao Zhou[3*], Qiangpeng Yang[3], Shilei Wen[3], and Kai Han[1†]

[1]The University of Hong Kong [2]Zhejiang University [3]ByteDance
huangxiaohu@connect.hku.hk, kaihanx@hku.hk

Abstract. In this paper, we present JoVA, a streamlined framework that unifies joint video-audio generation and editing. While existing methods often rely on fragmented, task-specific architectures or complex fusion mechanisms, JoVA employs native joint representation learning for direct video, audio, and text interaction in a dual-branch architecture. This design eliminates redundant alignment modules and effectively unifies diverse multimodal tasks within a single model. Furthermore, we utilize channel-wise conditioning for flexible image and video reference to avoid massive token expansion, alongside a mouth-area loss to enhance lip alignment. To fully empower and systematically evaluate this framework, we construct a comprehensive training corpus encompassing video-audio generation and editing datasets, and introduce unified benchmarks tailored for these multimodal tasks. Extensive experiments demonstrate that JoVA achieves state-of-the-art performance across benchmarks, establishing it as an extensible framework for versatile content creation. Project page: https://visual-ai.github.io/jova

1 Introduction

Recent years have witnessed a remarkable evolution in AI-driven content creation, transitioning from image synthesis to the more complex domain of video generation. This surge has catalyzed extensive research across various specialized directions. For instance, video generation models [6, 26, 33, 58, 71] have achieved high-fidelity visual synthesis; joint video-audio generation frameworks [22, 41, 59] have emerged to produce synchronized content; video editing techniques [4, 15, 36] enable fine-grained visual modifications; and lip-sync methods [35, 44, 47] focus on simulating realistic talking heads. However, these directions largely exist as separate streams of work.

To tackle these distinct tasks, researchers typically design task-specific model architectures or introduce complex multimodal fusion mechanisms, such as additional cross-attention layers [22, 41, 59]. This fragmentation limits model simplicity and extensibility. Recently, unified models [29, 55, 64] have demonstrated the ability to handle multiple video tasks within a single framework. Yet, they

* Equal contribution. †Corresponding author.

Fig. 1: Overview of the JoVA framework's capabilities. JoVA seamlessly unifies various multimodal tasks within a single model, including text-to-video-audio generation, reference-image-conditioned generation, video editing, and joint video-audio editing. By utilizing a simplified architecture, it ensures high-quality, synchronized, and flexible multimodal content creation.

lack the capability to output audio, rendering the generated video content less vivid and immersive. Therefore, developing a unified model capable of seamlessly handling diverse multimodal tasks—spanning both generation and editing across video and audio modalities—is a highly promising yet challenging frontier.

To bridge this gap, we present JoVA, a unified framework capable of joint video-audio generation and editing within a single model (Fig. 1). To realize this architecture, we integrate pretrained video and audio models, streamlining the overall design by eliminating redundant multimodal fusion modules. In their place, we use a native joint self-attention mechanism that concatenates video, audio, and text tokens, enabling direct and seamless multimodal interaction. This simplified configuration preserves the core philosophy of transformer architectures and guarantees exceptional task extensibility. To handle visual reference conditions in editing tasks, we adopt a channel-wise conditioning approach, which flexibly incorporates spatial inputs without drastically expanding the token sequence length. Furthermore, to address the challenge of lip-speech synchronization—where the mouth region occupies a tiny fraction of the frame ($\sim$2.3%)—we incorporate an area-specific loss during training to ensure precise alignment without adding architectural complexity.

To endow the model with versatile capabilities, we construct a large-scale training corpus of $\sim$3M diverse pairs, encompassing natural video-audio scenes, human talking videos, and complex joint editing data. For assessment, we intro-

duce two benchmarks, JoVABench-Gen and JoVABench-Edit, and further evaluate our model on the public Verse-Bench [59]. Extensive experiments demonstrate that our model consistently achieves leading performance across all tasks. Please refer to the supplementary material for the generated video results.

In summary, our main contributions are as follows:

- We propose JoVA, a unified architecture for joint video-audio modeling. By processing modalities through a unified representation, JoVA eliminates the reliance on complex cross-modal fusion modules and fragmented designs, establishing a highly extensible foundation for multimodal content creation.
- To empower this single architecture to handle a diverse spectrum of generation and editing scenarios, we construct a multi-domain training corpus of ∼3M high-quality pairs. The introduced automated data synthesis pipeline is designed to scale up the generation and editing of data.
- Extensive experiments on two introduced benchmarks (JoVABench-Gen and JoVABench-Edit) and Verse-Bench demonstrate that our single JoVA model achieves superior performance across diverse generation and editing tasks.

2 Related Work

2.1 Joint Video-Audio Generation

Recent advances in video generation have been driven by diffusion models transitioning from UNet-based architectures [49,50] to transformer-based designs [46], exemplified by Sora [8,45]. This shift has spurred high-fidelity models like Wan [1, 58], Goku [10], Waver [74], CogVideoX [71], and HunyuanVideo [33]. Beyond visual fidelity, recent post-training methods such as VGGRPO [3] introduce latent 4D geometry rewards to improve camera stability and world consistency in generated videos. Building upon this visual foundation, cross-modal generation largely diverged into unidirectional streams. Audio-driven video methods [11, 18, 19, 37] typically rely on cascaded pipelines to animate talking heads from speech. Conversely, video-to-audio (V2A) approaches [12, 23, 42, 53, 60, 75] focus on generating ambient sounds for silent videos. However, these decoupled approaches often require auxiliary synchronization modules, increasing system complexity and limiting semantic coherence. To achieve cohesive multimodal synthesis, various joint video-audio generation frameworks [28, 40, 51, 61, 67, 76] have emerged. Despite the significant progress marked by large-scale models like UniVerse-1 [59], Ovi [41], UniAVGen [73], and LTX-2 [22], most existing approaches adopt dual-branch architectures that process modalities separately and fuse them through explicit cross-attention. Such newly introduced alignment layers are typically trained on much smaller multimodal corpora than the billion-scale video backbones, which can disturb pretrained visual priors. In contrast, JoVA abandons complex fusion modules for native joint self-attention, reusing pretrained transformer parameters to process all modalities equally for precise alignment and architectural elegance.

2.2 Joint Video-Audio Editing

Instruction-driven video editing has advanced rapidly, with existing methods injecting conditions via cross-attention [55], channel or sequence concatenation [15, 24, 36, 64], or parameter-efficient fine-tuning [43]. Yet handling complex multi-task instructions in a single model remains difficult, and preserving audio-video synchronization adds further complexity. Recent works [17, 38, 77] employ training-free latent inversion or intricate pipelines with mask refiners and feedback agents. Similarly, precise lip-sync methods [35, 44, 47] are typically restricted to editing only the mouth region and heavily depend on external masks for guidance. A major drawback of these approaches is their fragmented nature: they often require stitching together independent models and auxiliary inputs, severely complicating the pipeline. In contrast, JoVA resolves these limitations through a natively unified architecture. It utilizes joint self-attention for multimodal interaction and channel-wise conditioning to avoid drastic sequence expansion, while a simple area-specific loss ensures accurate lip-sync.

3 Method

3.1 Preliminary

Our framework builds upon the Waver [74] video generation model, which employs a transformer-based diffusion architecture. For video encoding, we adopt the Wan2.1 VAE [58] to compress visual features in both spatial and temporal dimensions. Text prompts and instructions are processed through a combination of T5-XXL [13] and Qwen2.5-32B [69] encoders to provide rich semantic conditioning.

Following the MM-DiT architecture [16], Waver processes multimodal inputs through a two-stage design: an initial multi-stream stage where modalities maintain separate parameters, followed by a single-stream stage where all modalities share the same parameters. Each stage consists of transformer blocks containing self-attention and feed-forward network (FFN) modules. For a given layer, the computation can be expressed as:

$$\mathbf{h}' = \text{SelfAttn}(\mathbf{h}) + \mathbf{h}, \quad \mathbf{h}'' = \text{FFN}(\mathbf{h}') + \mathbf{h}', \tag{1}$$

where $\mathbf{h} \in \mathbb{R}^{N \times D}$ denotes the input features with N tokens and dimension D.

For training, we employ the flow matching [39] framework, which provides a simple and effective approach for generative modeling. The training objective minimizes the difference between the predicted velocity field and the true flow velocity:

$$\mathcal{L}_{\text{video}} = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1,\mathbf{C}} \left\| v_\theta(t, \mathbf{C}, \mathbf{x}_t) - u(\mathbf{x}_t|\mathbf{x}_0, \mathbf{x}_1) \right\|^2, \tag{2}$$

where $t \sim \mathcal{U}[0,1]$, $\mathbf{x}_0 \sim \mathcal{N}(0, \mathbf{I})$ is sampled noise, $\mathbf{x}_1 \sim q(\mathbf{x}_1, \mathbf{C})$ represents training data conditioned on $\mathbf{C}$ (text and other modalities), $\mathbf{x}_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0$ defines the interpolation path, and $u(\mathbf{x}_t|\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_1 - \mathbf{x}_0$ denotes the corresponding flow velocity.

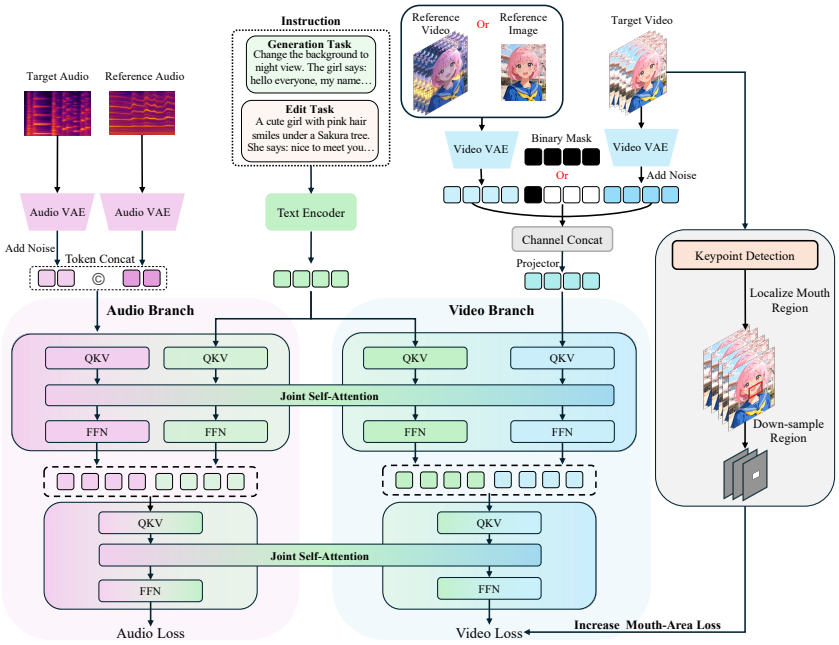

Fig. 2: Overall architecture of the proposed JoVA framework. Our model natively unifies video and audio generation and editing through a joint self-attention mechanism. Complex multi-task instructions are processed via a text encoder to guide the generation. For visual conditioning, we employ a channel-wise concatenation strategy with a binary mask to flexibly accommodate reference videos or images. Conversely, reference audio tokens are encoded and integrated into the audio branch by token concatenation. A localized mouth-area loss is applied to the video output to ensure lip-speech synchronization during training, which is exclusively active for human speech scenarios and does not affect the generation of natural scenes or ambient sounds.

3.2 Unified Video-Audio Generation and Editing

Joint Video-Audio Architecture We propose a unified framework designed to natively handle both generation and editing tasks across video and audio modalities. As illustrated in Fig. 2, our architecture comprises a visual branch, an audio branch, and a shared text encoder that processes complex multi-task instructions. Instead of relying on cascaded models or complex external fusion networks, we seamlessly integrate these modalities into a shared transformer backbone.

To establish robust audio generation capabilities within this single model, we initialize the audio diffusion model by duplicating the pretrained parameters of our video backbone. For feature extraction, we employ the MMAudio VAE [12] to process audio spectrograms, converting raw audio signals into compact latent representations. Within each transformer block, tokens from the video input, the audio input, and the instruction text are processed through a shared joint

self-attention mechanism:

$$[\mathbf{h}'_v; \mathbf{h}'_a; \mathbf{h}'_t] = \text{JointAttn}([\mathbf{h}_v; \mathbf{h}_a; \mathbf{h}_t]), \tag{3}$$

where $\mathbf{h}_t$ contains text features for both the video and audio branches. Unlike prior methods that rely on separate cross-attention layers, our joint self-attention mechanism allows video, audio, and text tokens to interact directly in a unified space. It also avoids adding newly initialized alignment modules whose training data scale is much smaller than that of the pretrained backbone. Furthermore, to ensure precise temporal alignment between the video and audio streams, we adopt the temporal-aligned RoPE (Rotary Position Embedding) from MMAudio [12], which synchronizes the positional encodings of both modalities along the temporal dimension.

Unified Conditioning. To accommodate editing instructions, we introduce modality-specific conditional inputs. The integration strategies are carefully tailored to the distinct sequence characteristics of video and audio modalities.

For the audio branch, alongside the noisy target audio latent, we introduce a reference audio latent encoded via the Audio VAE to provide acoustic context. Because the sequence length of audio latents is extremely compact—typically accounting for less than 1% of the video token length—we can directly employ a token concatenation strategy to integrate the reference audio. This approach effectively provides the necessary context without incurring significant computational overhead.

Conversely, for the visual branch, directly appending reference video frames via token concatenation would lead to a massive increase in sequence length, resulting in prohibitive computational costs. Therefore, we employ a channel-wise conditioning strategy guided by a binary mask. Specifically, the reference visual input is encoded by the Video VAE and concatenated with the noisy target video latent and a corresponding mask along the channel dimension. The mask is set to 1 for frames where the reference is provided, and 0 otherwise (in which case the reference channels are filled with default padding). After the channels are concatenated, a projector layer is utilized to fuse their information. This design not only avoids the token explosion problem but also natively unifies image-to-video generation and video editing (flexibly accommodating reference images as conditions) without requiring architectural modifications.

Training Objectives and Mouth-Area Loss Our unified model is optimized using flow matching objectives. To ensure the model learns a wide variety of acoustic patterns and human speech characteristics, the audio branch is trained on a comprehensive mixture of large-scale text-to-audio and text-to-speech datasets. The audio flow matching loss is defined as:

$$\mathcal{L}_{\text{audio}} = \mathbb{E}_{t,\mathbf{a}_0,\mathbf{a}_1,\mathbf{C}} \left\| v_\theta(t, \mathbf{C}, \mathbf{a}_t) - u(\mathbf{a}_t | \mathbf{a}_0, \mathbf{a}_1) \right\|^2, \tag{4}$$

where $\mathbf{a}_t$ represents the audio latent at time t, and $\mathbf{C}$ denotes the conditioning information.

While the unified architecture effectively handles general video-audio alignment, achieving fine-grained lip-speech synchronization for human-centric content requires targeted supervision. To address this, we introduce a localized

strategy that emphasizes the mouth region during training. We first apply facial keypoint detection to the original video frames to localize the mouth region and extract a bounding box. This bounding box is then mapped to the VAE latent space through proportional spatial scaling and a sliding-window temporal downsampling to match the latent dimensions, yielding a precise mouth region mask $\mathcal{M}$. More implementation details are given in the supplementary material.

Our final training objective combines the general flow matching losses with this localized supervision:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{video}} + \mathcal{L}_{\text{audio}} + \lambda \mathcal{L}_{\text{mouth}}, \tag{5}$$

where $\mathcal{L}_{\text{video}}$ is the standard video flow matching loss. The term $\mathcal{L}_{\text{mouth}}$ applies the flow matching objective strictly restricted to the mouth region of the video latent:

$$\mathcal{L}_{\text{mouth}} = \mathbb{E}_{t,\mathbf{v}_0,\mathbf{v}_1,\mathbf{C}} \left\| v_\theta(t, \mathbf{C}, \mathbf{v}_t)_{\mathcal{M}} - u(\mathbf{v}_t|\mathbf{v}_0, \mathbf{v}_1)_{\mathcal{M}} \right\|^2, \tag{6}$$

where $\mathbf{v}_t$ denotes the video latent at time t, and λ is a weighting coefficient. Crucially, this mouth-area loss is exclusively applied to data containing human speech. For natural scenes or general ambient sounds, we set λ to 0, ensuring that the generation of non-human content remains completely unaffected.

3.3 Construction of Training Data

To support unified video-audio generation and editing across diverse scenarios, we construct a large-scale, high-quality dataset comprising approximately 3 million (3M) video-audio-text pairs. Fig. 4 summarizes the dataset composition, while Fig. 3 illustrates the two primary construction pipelines tailored for generation and editing tasks, respectively.

Data Construction for Video-Audio Generation. As shown in Fig. 3(a), we begin with a large corpus of raw video clips. To ensure data quality, these clips first pass through a rigorous Quality Assessment module, utilizing both Audio and Video Assessment Models to filter out low-quality samples based on predefined thresholds. The retained high-quality videos are then categorized into two streams: videos with speech and videos without speech. For videos containing speech, we apply a Lip-sync Filtering module to ensure tight audio-visual alignment, which is crucial for talking-head generation. For videos without speech, a Speech Filter is applied to guarantee the purity of ambient sounds. Subsequently, we employ a multimodal annotation pipeline to generate comprehensive text labels. We utilize Tarsier [72] for detailed video captioning and Audio-flamingo [21] for audio captioning. For the speech subset, Whisper [48] is additionally used for automatic speech recognition (ASR) to extract precise transcripts. This tri-level annotation enriches each sample with structured semantic information from complementary visual, auditory, and linguistic perspectives.

Data Construction for Video-Audio Editing. To train the model for video-audio editing tasks, we design an automated data synthesis pipeline (Fig. 3b) to construct paired source-target editing data. Inspired by the visual editing paradigm of OpenVE [24], we adapt and extend this approach to the multimodal

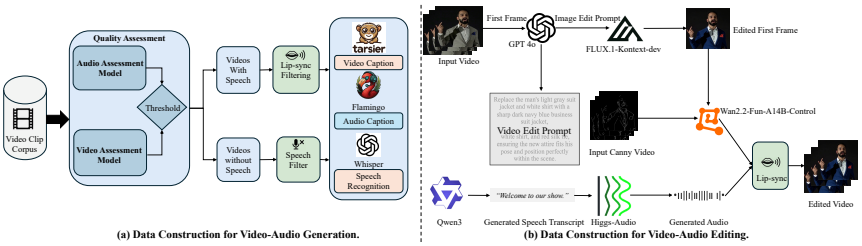

Fig. 3: Overview of our training data construction pipelines. (a) **Video-Audio Generation Data:** Raw videos undergo rigorous quality assessment and filtering, followed by comprehensive multimodal annotation including video/audio captioning and speech recognition. (b) **Video-Audio Editing Data:** A workflow to synthesize paired editing data, utilizing foundation models for visual editing, audio generation, and lip-sync alignment.

domain. Given an input video, we extract its first frame and utilize GPT-4o to generate an Image Edit Prompt. This prompt guides FLUX.1-Kontext-dev [5] to produce an Edited First Frame. Concurrently, GPT-4o [27] generates a Video Edit Prompt detailing the desired temporal changes. We then extract the Canny edge map from the input video to serve as structural guidance. The Edited First Frame, Input Canny Video, and Video Edit Prompt are fed into Wan2.2-Fun-A14B-Control [2] to generate the visual component of the edited video. Building upon this visual foundation, we further construct the corresponding audio track. We use Qwen3 [68] to generate a speech transcript, which is then synthesized into high-quality audio using Higgs-Audio [7]. Finally, a Lip-sync module aligns the generated audio with the visual output, resulting in the final Edited Video. This pipeline allows us to efficiently scale up high-quality, paired multimodal editing data.

3.4 Dataset Distribution and Details

To train JoVA effectively for both generation and editing tasks, we construct a large-scale, diverse, and highly aligned video-audio-text dataset. As shown in Fig. 4, the dataset encompasses approximately 3 million high-quality samples and covers both generation-oriented and editing-oriented data.

Overview of the Training Dataset. The overall dataset is broadly categorized into Generation Data and Edit Data, which are further divided into three main subsets:

- **Text-Video-Audio General Scenes (without Speech):** This subset comprises ∼650K samples (21.6% of the overall dataset). Sourced from open-source datasets including InternVID [63], AudioCaps [31], and VGGSound [9], it is strictly cleaned and aligned to help the model learn correspondences between visual elements and natural environmental sounds.
- **Text-Video-Audio Human Speech Scenes:** This subset comprises ∼1M samples (33.2% of the overall dataset). Collected and annotated by our team, it focuses on strong synchronization among lip movements, facial expressions, and speech in human talking scenarios.

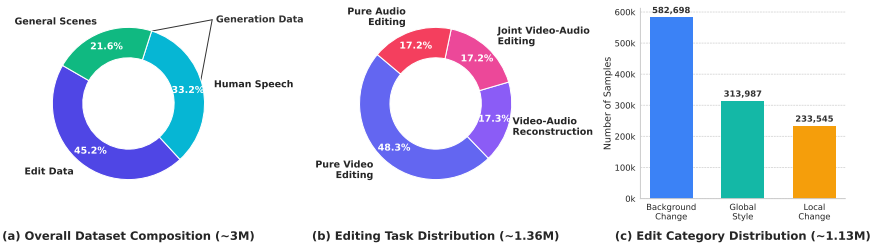

Fig. 4: Detailed statistics of our video-audio-text dataset. (a) Overall dataset composition (∼3M), highlighting the proportion of Edit Data (45.2%) versus Generation Data, including General Scenes (21.6%) and Human Speech (33.2%). (b) Distribution of the ∼1.36M editing task samples across four task types. (c) Distribution of the ∼1.13M edit category samples, excluding reconstruction tasks.

– **Video-Audio Edit Data:** This subset comprises ∼1.36M samples (45.2% of the overall dataset). Derived from the human speech scenes, it is constructed using our automated pipeline to generate rich editing instructions and source-target pairs, establishing the foundation for JoVA's editing capabilities.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate our method on three benchmarks. First, we introduce JoVABench-Gen and JoVABench-Edit, two curated evaluation sets specifically designed to assess video-audio generation and editing capabilities, respectively. We emphasize that all 200 samples in these benchmarks undergo a strict filtering process to ensure there is absolutely no overlap with our training data. Second, to evaluate the broader generalization capabilities of our model, we employ Verse-Bench [59] as an out-of-domain dataset. It contains 600 image-text prompt pairs spanning a wide range of audio categories, including human speech, animal vocalizations, instrumental music, and natural ambient sounds.

Evaluation Metrics. Following the evaluation protocol of UniVerse-1 [59], we assess our model across multiple dimensions: joint video-audio quality, text-to-speech accuracy, audio generation quality, and video generation fidelity. For video-audio alignment, we measure lip-sync accuracy using SyncNet [14] to calculate confidence scores (LSE-C) and distances (LSE-D) only on human-speech samples; general-scene videos are strictly excluded because these metrics target visible mouth motion. For general-scene audio-video synchronization, we additionally report DeSync, where lower values indicate better temporal alignment. Text-to-speech quality is evaluated through Word Error Rate (WER) by transcribing generated audio with Whisper-large-v3 [48]. Video generation fidelity is comprehensively assessed using Inception Score (IS) [52], computed on the generated video frames rather than the audio, for visual quality and diversity, Motion Score (MS) calculated from RAFT-detected [56] optical flow, Aesthetic Score (AS) averaging MANIQA [70] fidelity, aesthetic-predictor-v2-5

quality, and Musiq [30] scores, ID consistency measuring DINOV3 [54] feature similarity between reference images and generated frames, and frame consistency (FC) measuring CLIP similarity between consecutive frames. Additionally, we evaluate text-video semantic consistency using ImageBind [20] Text-Video Alignment (IB-TV). For audio generation, we evaluate distributional similarity using Fréchet Distance (FD) and Kullback-Leibler (KL) divergence on PANNs [32] and PaSST [34] features, semantic consistency via LAION-CLAP [66] scores, and AudioBox-Aesthetics [57] measures for Production Quality (PQ), Production Complexity (PC), Content Enjoyment (CE), and Content Usefulness (CU).

4.2 Implementation Details

Unless stated, our framework builds upon the Waver [74] 12B model as the base architecture. We adopt a two-stage training strategy: in the first stage, we train the audio branch in isolation on collected text-to-audio and text-to-speech data using a batch size of 256 and a fixed learning rate of 1×10^{-4} for 80K steps. In the second stage, we perform joint video-audio generation and editing training with a batch size of 32 and a learning rate of 3×10^{-5} for 100K steps to learn multimodal interaction. During inference, we apply classifier-free guidance [25] with a weight of 10.0 for both video and audio generation and editing to enhance generation quality.

4.3 Comparison with State-of-the-Art Models

On joint video-audio generation benchmarks (JoVABench-Gen and Verse-Bench), we compare our method against two categories of approaches: (1) audio-driven generation methods, where we use ground-truth audio to drive video synthesis for fair comparison of conditional generation capability, including FantasyTalking [62] and Wan-S2V [19]; (2) joint video-audio generation methods, including the recent UniVerse-1 [59], JavisDiT [40], Ovi [41], LTX-2 [22], and UniAVGen [73] where available.

JoVABench-Gen. Table 1 shows that JoVA achieves the best overall performance on JoVABench-Gen. It reaches the highest LSE-C (6.70), surpassing Ovi (6.41), LTX-2 (6.21), UniAVGen (5.87), UniVerse-1 (1.62), and the audio-driven Wan-S2V (6.43), which validates the effect of mouth-area supervision. JoVA also obtains strong speech, audio, and video quality, including the best FD (0.67), CE (5.40), PQ (6.49), MS (0.97), and AS (0.48). On the shared metrics from LTX-2 and UniAVGen, JoVA leads on LSE-C, FD, CE, PQ, MS, AS, and ID while remaining competitive on WER, KL, CU, and PC.

JoVABench-Edit. Table 2 evaluates JoVA on JoVABench-Edit. Since no directly comparable open-source baseline exists for joint video-audio editing, we construct cascaded pipelines that combine video editing models with Diff2Lip or Wav2Lip using ground-truth audio. These baselines therefore solve a strictly easier conditional task, while JoVA jointly synthesizes both video and audio. Despite this advantage, JoVA achieves the best lip-sync accuracy (LSE-C 5.88, LSE-D 1.66) and the strongest video fidelity (IS 3.21, AS 0.47).

Table 1: Performance comparison on JoVABench-Gen. The results of other methods are reproduced using their official code. The audio inputs for audio-driven models are from ground-truth audio. Missing reported metrics are denoted by "-". ↑ / ↓ indicate higher / lower is better.

Method	Audio-Video	TTS	Audio							Video		
	LSE-C↑	WER↓	FD↓	KL↓	CS↑	CE↑	CU↑	PC↓	PQ↑	MS↑	AS↑	ID↑
Audio-driven Generation												
Fantasy-Talking [62]	3.10	-	-	-	-	-	-	-	-	0.22	0.44	**0.87**
Wan-S2V [19]	6.43	-	-	-	-	-	-	-	-	0.82	0.44	0.72
Joint Video-Audio Generation												
UniVerse-1 [59]	1.62	0.37	1.04	0.83	0.16	3.68	3.90	2.12	4.39	0.43	0.42	0.82
JavisDiT [40]	1.04	1.08	1.15	0.64	**0.41**	3.36	3.53	2.31	4.76	0.20	0.44	0.30
Ovi [41]	6.41	0.23	0.75	0.66	0.30	5.00	5.67	1.75	5.77	0.94	0.41	0.75
LTX-2 [22]	6.21	**0.15**	0.80	0.64	0.32	5.06	5.53	1.81	5.64	0.92	0.44	0.74
UniAVGen [73]	5.87	0.17	0.70	**0.62**	0.34	5.32	**6.10**	**1.62**	6.30	0.90	0.44	0.75
JoVA	**6.70**	0.19	**0.67**	0.64	0.33	**5.40**	6.03	1.69	**6.49**	**0.97**	**0.48**	0.78

JoVA does not rank first in every single metric, but it provides the best overall trade-off. Its MS of 0.66 is close to OmniVideo+Diff2Lip (0.71), while avoiding that pipeline's lower lip-sync score (LSE-C 4.07). It also maintains competitive IB-TV (0.21) and FC (0.98) without sacrificing video quality (IS/AS), showing the advantage of avoiding error accumulation in two-stage pipelines.

Verse-Bench. On Verse-Bench (Table 3), JoVA remains strong across diverse scenarios. It achieves the lowest WER (0.11), competitive LSE-C (6.65), the best DeSync (0.16), FD (0.82), KL (1.39), CS (0.29), CE (4.47), MS (0.75), and AS (0.49), while preserving identity consistency (0.86).

4.4 Diagnostic Study

Effectiveness of Mouth-Area Loss. We first investigate the impact of the mouth-area loss weight λ in Eq. (5). As shown in Table 4, increasing λ from 0.0 dramatically improves lip-sync accuracy (LSE-C from 1.39 to 6.53), demonstrating the critical role of targeted supervision on the mouth region. Without mouth-area loss ($\lambda = 0.0$), the model does not fully learn fine-grained lip-speech alignment through joint attention alone. The optimal performance is achieved at $\lambda = 5.0$ (LSE-C: 6.70), where the model effectively balances lip-sync precision with overall generation quality. Further increasing λ to 8.0 leads to marginal degradation (LSE-C: 6.65). Notably, audio quality metrics (WER, KL, PQ) and AS remain stable across different λ values, confirming that our mouth-area supervision strategy enhances lip-sync without sacrificing general audio-visual quality. The substantial improvement from $\lambda = 0.0$ to $\lambda = 2.0$ (LSE-C: 1.39 $\rightarrow$ 6.53) highlights the necessity of explicit mouth-region guidance for achieving precise synchronization.

Comparison of Joint Self-Attention and Cross-Attention. To validate our joint self-attention design, we compare it with three cross-attention variants under identical capacity and training data. As shown in Fig. 5, JoVA processes

Table 2: Performance comparison on JoVABench-Edit. As there are no directly comparable open-source baselines, we combine video editing models with lip-sync models (Diff2Lip [44] and Wav2Lip [47]) for comparison, which use ground-truth audio as inputs.

Method	LSE-C↑	LSE-D↓	FC↑	IB-TV↑	IS↑	AS↑	MS↑
Combined with Diff2Lip							
Ditto [4] + Diff2Lip [44]	2.44	5.30	0.98	**0.23**	2.13	0.34	0.35
ICVE [36] + Diff2Lip [44]	2.45	5.35	**0.99**	0.18	1.72	0.36	0.30
InsViE [65] + Diff2Lip [44]	2.38	5.72	**0.99**	0.21	1.02	0.39	0.28
Lucy [15] + Diff2Lip [44]	2.34	5.83	0.98	0.18	3.01	0.28	0.30
OmniVideo [55] + Diff2Lip [44]	4.07	5.50	0.98	0.19	3.03	0.35	**0.71**
Combined with Wav2Lip							
Ditto [4] + Wav2Lip [47]	5.85	2.83	0.98	**0.23**	2.33	0.42	0.46
ICVE [36] + Wav2Lip [47]	5.54	2.50	0.98	0.15	1.55	0.41	0.32
InsViE [65] + Wav2Lip [47]	5.16	4.40	0.97	0.18	1.36	0.37	0.24
Lucy [15] + Wav2Lip [47]	5.37	3.24	0.98	0.19	3.01	0.34	0.53
OmniVideo [55] + Wav2Lip [47]	4.86	1.75	0.97	0.18	2.27	0.37	0.63
JoVA	**5.88**	**1.66**	0.98	0.21	**3.21**	**0.47**	0.66

Table 3: Performance comparison on Verse-Bench. The results of Ovi and JavisDiT are reproduced using their official code. Missing reported metrics are denoted by "-". ↑ / ↓ indicate higher / lower is better.

Method	Audio-Video		TTS	Audio							Video		
	LSE-C↑	DeSync↓	WER↓	FD↓	KL↓	CS↑	CE↑	CU↑	PC↓	PQ↑	MS↑	AS↑	ID↑
Audio-driven Generation													
FantasyTalking [62]	2.68	-	-	-	-	-	-	-	-	-	0.07	0.42	0.87
Wan-S2V [19]	6.49	-	-	-	-	-	-	-	-	-	0.17	0.46	0.89
Joint Video-Audio Generation													
UniVerse-1 [59]	1.62	0.23	0.18	1.25	2.70	0.16	3.53	4.61	**2.49**	5.20	0.20	0.47	**0.89**
JavisDiT [40]	0.85	0.27	1.00	0.99	1.75	0.19	3.48	5.03	2.49	5.49	0.27	0.42	0.40
Ovi [41]	6.61	0.49	0.12	0.97	2.23	0.20	3.95	5.60	2.09	6.20	0.55	0.47	0.86
JoVA	**6.65**	**0.16**	**0.11**	**0.82**	**1.39**	**0.29**	**4.47**	**5.94**	2.47	**6.23**	**0.75**	**0.49**	0.86

video, audio, and text tokens together in unified self-attention, enabling bidirectional multimodal interaction while reusing pretrained transformer parameters. In contrast, the sequential Ovi-style variant alternates cross-modal interaction, while the two parallel variants compute cross-modal attention alongside self-attention and merge the results either with or without an additional linear adaptation layer.

Tab. 5 shows that joint self-attention achieves the best LSE-C (6.70), WER (0.19), PQ (6.49), ID (0.78), and matching-best AS (0.48). The low LSE-C of the parallel cross-attention variant with linear adaptation (1.59) suggests that the added projection can disrupt cross-modal alignment. These results validate that unified multimodal processing is essential for high-quality, synchronized audio-visual generation without introducing additional parameters.

4.5 Qualitative Results

Fig. 6 presents qualitative comparisons of video-audio generation quality. We visualize generated frames alongside their audio waveforms and automatic speech recognition (ASR) results, with recognition errors underlined in red. UniVerse-1

Table 4: Effect of mouth-area loss weight λ on JoVABench-Gen. ↑ / ↓ indicate higher / lower is better.

λ	LSE-C↑	WER↓	FD↓	KL↓	PQ↑	AS↑
0.0	1.39	0.17	0.67	0.63	6.44	0.48
2.0	6.53	0.18	0.67	0.62	6.47	0.46
5.0	6.70	0.19	0.67	0.64	6.49	0.48
8.0	6.65	0.18	0.68	0.63	6.40	0.47

Table 5: Comparison of audio-video fusion strategies on JoVABench-Gen. All variants share the same backbone capacity, training data, and schedule; only the cross-modal interaction mechanism differs. ↑ / ↓ indicate higher / lower is better.

Strategy	LSE-C↑	WER↓	PQ↑	AS↑	ID↑
Seq. Cross-Attn. (Ovi-style)	6.51	0.24	6.25	0.48	0.75
Parallel Cross-Attn. (w/ linear)	1.59	0.29	6.31	0.46	0.74
Parallel Cross-Attn. (w/o linear)	6.57	0.28	6.14	0.47	0.76
Joint Self-Attn. (Ours)	**6.70**	**0.19**	**6.49**	**0.48**	**0.78**

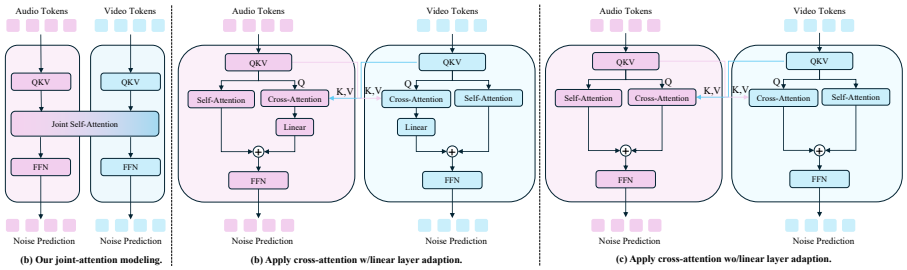

Fig. 5: Comparison of different audio-video fusion strategies. (a) Joint self-attention processes video, audio, and text tokens together in one layer. (b,c) Cross-attention variants exchange information through explicit audio-video attention paths, while text conditioning remains inside the self-attention layers.

produces incomplete speech with missing words, indicating weaker audio generation quality. Ovi shows temporal consistency issues, including sudden freezing artifacts and mispronunciations. In contrast, JoVA generates smooth, natural speech with accurate word pronunciation and maintains precise lip-speech synchronization throughout the sequence. The aligned audio waveforms and correct ASR transcriptions demonstrate JoVA's superior capability in joint video-audio generation with proper cross-modal alignment.

Fig. 7 further compares video-audio editing. JoVA better follows background, style, and local edit instructions while preserving unedited regions, whereas baselines tend to introduce identity shifts, entangle unrelated attributes, or produce incomplete edits under the same audio-conditioned setting. For background replacement, JoVA changes the scene while preserving the subject appearance; for style conversion, it restores photorealism without corrupting temporal consistency; for local replacement, it edits target attributes without changing the global layout. These cases show that joint video-audio conditioning provides stronger instruction following than cascaded editing baselines. More visualizations are given in the supplementary materials.

5 Conclusion

In this paper, we introduce JoVA, an extensible framework that unifies joint video-audio generation and editing. Native joint self-attention eliminates com-

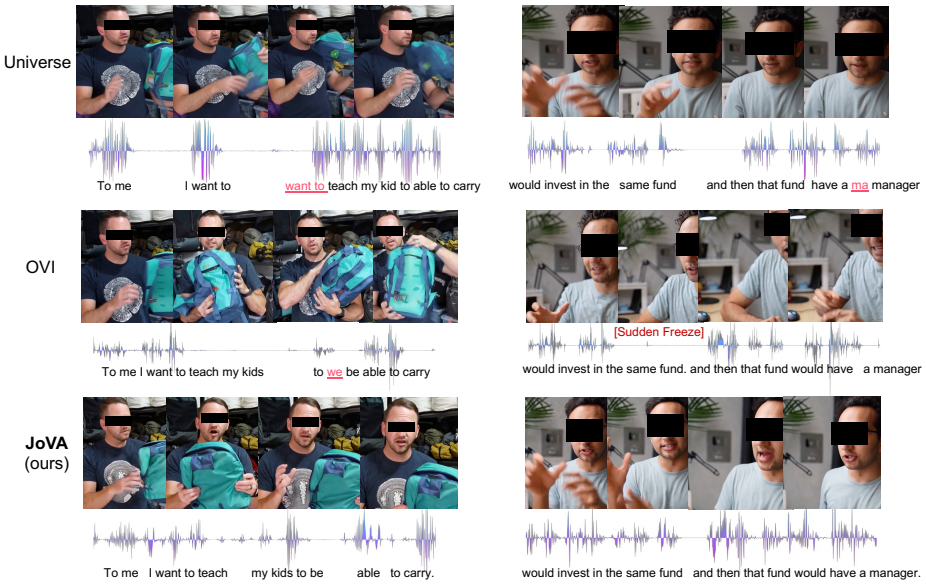

Fig. 6: Qualitative comparison of speech generation and lip-sync quality. Generated video frames with audio waveforms and ASR transcriptions. Red underlines indicate speech recognition errors. UniVerse-1 omits words, Ovi shows freezing artifacts and mispronunciations, while JoVA produces accurate speech with precise lip-sync alignment.

plex task-specific fusion modules and enables direct interaction among video, audio, and text, while channel-wise conditioning supports flexible visual editing without prohibitive token growth. A simple mouth-area loss further improves lip-speech synchronization for human-centric content.

To train and evaluate JoVA, we curate a comprehensive multi-domain corpus and benchmarks. Experiments show state-of-the-art performance across generation and editing tasks, especially in multimodal alignment and generation quality. We also observe several limitations: very fast hand motion can still induce finger artifacts in generation, and instruction-based editing does not always preserve text in text-rich backgrounds. In addition, our audio-video editing currently focuses on human-speech scenes, as paired audio-video editing data for general scenes remains difficult to obtain; this is a data limitation rather than an architectural one, and our framework can be extended to general-scene editing once suitable data becomes available. Future work will address these issues by expanding training diversity and broadening JoVA's coverage of high-motion scenes, text-rich backgrounds, and general-scene audio-video editing. JoVA establishes a strong foundation for unified multimodal content creation.

Acknowledgements. This work is supported by Hong Kong Research Grants Council – General Research Fund (Grant No. 17211024), Hong Kong Innovation and Technology Commission – Innovation and Technology Fund (Grant No. ITS/488/24FP), and HKU Seed Fund for PI Research.

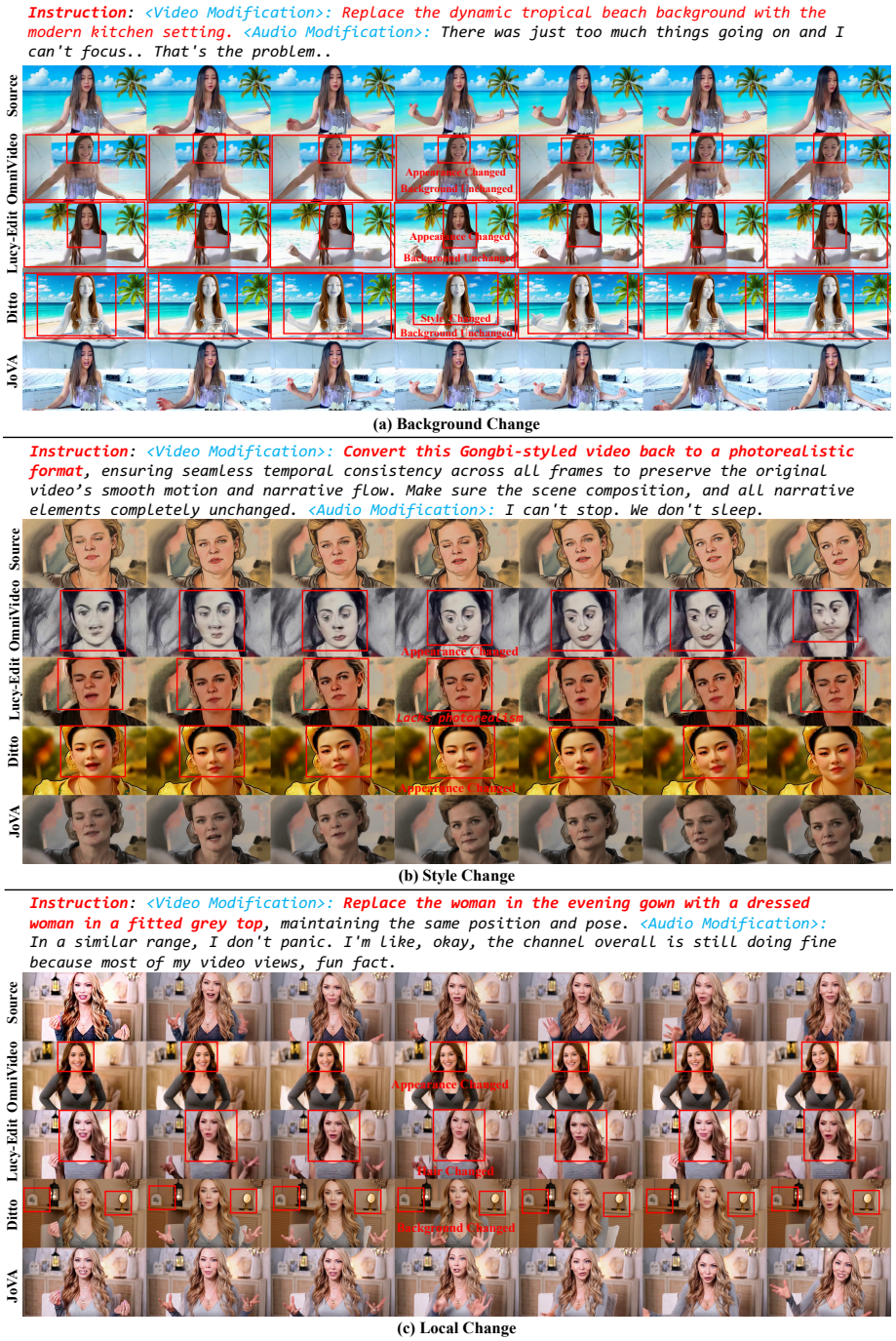

Fig. 7: Qualitative comparison of video-audio editing. We compare JoVA against cascaded baselines (OmniVideo + Diff2Lip and Ditto + Wav2Lip, with ground-truth audio) across background change, style change, and local change. JoVA better follows the editing instructions while preserving unedited regions, whereas the baseline pipelines often introduce identity shifts, attribute entanglement, or incomplete edits. Pose differences are expected because the edited subjects are driven by modified audio inputs.

References

1. Alibaba Cloud: Wan2.5. https://wan.video/ (2025), accessed: 2026-06-30
2. Alibaba PAI: Wan2.2-Fun-A14B-Control. https://huggingface.co/alibaba-pai/Wan2.2-Fun-A14B-Control (2025), huggingFace release. Accessed: 2026-06-30
3. An, Z., Kupyn, O., Uscidda, T., Colaco, A., Ahuja, K., Belongie, S., Gonzalez-Franco, M., Gazulla, M.T.: VGGRPO: Towards world-consistent video generation with 4D latent reward. arXiv preprint arXiv:2603.26599 (2026)
4. Bai, Q., Wang, Q., Ouyang, H., Yu, Y., Wang, H., Wang, W., Cheng, K.L., Ma, S., Zeng, Y., Liu, Z., et al.: Scaling instruction-based video editing with a high-quality synthetic dataset. arXiv preprint arXiv:2510.15742 (2025)
5. Black Forest Labs, Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: FLUX.1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
6. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
7. Boson AI: Higgs-Audio V2: Redefining expressiveness in audio generation. https://github.com/boson-ai/higgs-audio (2025), accessed: 2026-06-30
8. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators. https://openai.com/index/sora/ (2024), OpenAI Technical Report. Accessed: 2026-06-30
9. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP (2020)
10. Chen, S., Ge, C., Zhang, Y., Zhang, Y., Zhu, F., Yang, H., Hao, H., Wu, H., Lai, Z., Hu, Y., et al.: Goku: Flow based video generative foundation models. In: CVPR. pp. 23516–23527 (2025)
11. Chen, Y., Liang, S., Zhou, Z., Huang, Z., Ma, Y., Tang, J., Lin, Q., Zhou, Y., Lu, Q.: HunyuanVideo-Avatar: High-fidelity audio-driven human animation for multiple characters. arXiv preprint arXiv:2505.20156 (2025)
12. Cheng, H.K., Ishii, M., Hayakawa, A., Shibuya, T., Schwing, A., Mitsufuji, Y.: MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In: CVPR (2025)
13. Chung, H.W., Constant, N., Garcia, X., Roberts, A., Tay, Y., Narang, S., Firat, O.: Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. arXiv preprint arXiv:2304.09151 (2023)
14. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: ACCV (2016)
15. DecartAI Team: Lucy edit: Open-weight text-guided video editing. https://huggingface.co/decart-ai/Lucy-Edit-Dev (2025), accessed: 2026-06-30
16. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
17. Fu, Y., Si, R., Wang, H., Zhou, D., Sun, J., Luo, P., Hu, D., Zhang, H., Li, X.: Object-AVEdit: An object-level audio-visual editing model. arXiv preprint arXiv:2510.00050 (2025)

18. Gan, Q., Yang, R., Zhu, J., Xue, S., Hoi, S.: OmniAvatar: Efficient audio-driven avatar video generation with adaptive body animation. arXiv preprint arXiv:2506.18866 (2025)
19. Gao, X., Hu, L., Hu, S., Huang, M., Ji, C., Meng, D., Qi, J., Qiao, P., Shen, Z., Song, Y., et al.: Wan-S2V: Audio-driven cinematic video generation. arXiv preprint arXiv:2508.18621 (2025)
20. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: ImageBind: One embedding space to bind them all. In: CVPR. pp. 15180–15190 (2023)
21. Goel, A., Ghosh, S., Kim, J., Kumar, S., Kong, Z., Lee, S.g., Yang, C.H.H., Duraiswami, R., Manocha, D., Valle, R., et al.: Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. arXiv preprint arXiv:2507.08128 (2025)
22. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., et al.: LTX-2: Efficient joint audio-visual foundation model. arXiv preprint arXiv:2601.03233 (2026)
23. Haji-Ali, M., Menapace, W., Siarohin, A., Skorokhodov, I., Canberk, A., Lee, K.S., Ordonez, V., Tulyakov, S.: AV-Link: Temporally-aligned diffusion features for cross-modal audio-video generation. arXiv preprint arXiv:2412.15191 (2024)
24. He, H., Wang, J., Zhang, J., Xue, Z., Bu, X., Yang, Q., Wen, S., Xie, L.: OpenVE-3M: A large-scale high-quality dataset for instruction-guided video editing. arXiv preprint arXiv:2512.07826 (2025)
25. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
26. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
27. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: GPT-4o system card. arXiv preprint arXiv:2410.21276 (2024)
28. Ishii, M., Hayakawa, A., Shibuya, T., Mitsufuji, Y.: A simple but strong baseline for sounding video generation: Effective adaptation of audio and video diffusion models for joint generation. arXiv preprint arXiv:2409.17550 (2024)
29. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: ICCV. pp. 17191–17202 (2025)
30. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: ICCV (2021)
31. Kim, C.D., Kim, B., Lee, H., Kim, G.: Audiocaps: Generating captions for audios in the wild. In: ACL (2019)
32. Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D.: PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2020)
33. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: HunyuanVideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
34. Koutini, K., Schlüter, J., Eghbal-Zadeh, H., Widmer, G.: Efficient training of audio transformers with patchout. arXiv preprint arXiv:2110.05069 (2021)
35. Li, C., Zhang, C., Xu, W., Lin, J., Xie, J., Feng, W., Peng, B., Chen, C., Xing, W.: LatentSync: Taming audio-conditioned latent diffusion models for lip sync with SyncNet supervision. arXiv preprint arXiv:2412.09262 (2024)

36. Liao, X., Zeng, X., Song, Z., Fu, Z., Yu, G., Lin, G.: In-context learning with unpaired clips for instruction-based video editing. arXiv preprint arXiv:2510.14648 (2025)
37. Lin, G., Jiang, J., Yang, J., Zheng, Z., Liang, C.: Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. arXiv preprint arXiv:2502.01061 (2025)
38. Lin, Y.B., Lin, K., Yang, Z., Li, L., Wang, J., Lin, C.C., Wang, X., Bertasius, G., Wang, L.: Zero-shot audio-visual editing via cross-modal delta denoising. In: WACV (2026)
39. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
40. Liu, K., Li, W., Chen, L., Wu, S., Zheng, Y., Ji, J., Zhou, F., Jiang, R., Luo, J., Fei, H., et al.: JavisDiT: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. arXiv preprint arXiv:2503.23377 (2025)
41. Low, C., Wang, W., Katyal, C.: Ovi: Twin backbone cross-modal fusion for audio-video generation. arXiv preprint arXiv:2510.01284 (2025)
42. Luo, S., Yan, C., Hu, C., Zhao, H.: Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. NeurIPS (2023)
43. Mou, C., Sun, Q., Wu, Y., Zhang, P., Li, X., Ye, F., Zhao, S., He, Q.: InstructX: Towards unified visual editing with MLLM guidance. arXiv preprint arXiv:2510.08485 (2025)
44. Mukhopadhyay, S., Suri, S., Gadde, R.T., Shrivastava, A.: Diff2Lip: Audio conditioned diffusion models for lip-synchronization. In: WACV (2024)
45. OpenAI: Sora2. `https://openai.com/index/sora-2-system-card/` (2025), accessed: 2026-06-30
46. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4195–4205 (2023)
47. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: ACM MM (2020)
48. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: ICML (2023)
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
50. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
51. Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N.J., Jin, Q., Guo, B.: Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In: CVPR (2023)
52. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: NeurIPS (2016)
53. Shan, S., Li, Q., Cui, Y., Yang, M., Wang, Y., Yang, Q., Zhou, J., Zhong, Z.: HunyuanVideo-Foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. arXiv preprint arXiv:2508.16930 (2025)
54. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)
55. Tan, Z., Yang, H., Qin, L., Gong, J., Yang, M., Li, H.: Omni-video: Democratizing unified video understanding and generation. arXiv preprint arXiv:2507.06119 (2025)
56. Teed, Z., Deng, J.: RAFT: Recurrent all-pairs field transforms for optical flow. In: ECCV (2020)

57. Tjandra, A., Wu, Y.C., Guo, B., Hoffman, J., Ellis, B., Vyas, A., Shi, B., Chen, S., Le, M., Zacharov, N., et al.: Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. arXiv preprint arXiv:2502.05139 (2025)
58. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
59. Wang, D., Zuo, W., Li, A., Chen, L.H., Liao, X., Zhou, D., Yin, Z., Dai, X., Jiang, D., Yu, G.: UniVerse-1: Unified audio-video generation via stitching of experts. arXiv preprint arXiv:2509.06155 (2025)
60. Wang, J., Zeng, X., Qiang, C., Chen, R., Wang, S., Wang, L., Zhou, W., Cai, P., Zhao, J., Li, N., et al.: Kling-Foley: Multimodal diffusion transformer for high-quality video-to-audio generation. arXiv preprint arXiv:2506.19774 (2025)
61. Wang, K., Deng, S., Shi, J., Hatzinakos, D., Tian, Y.: AV-DiT: Taming image diffusion transformers for efficient joint audio and video generation. In: ACM MM (2025)
62. Wang, M., Wang, Q., Jiang, F., Fan, Y., Zhang, Y., Qi, Y., Zhao, K., Xu, M.: Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. In: ACM MM (2025)
63. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. arXiv preprint arXiv:2307.06942 (2023)
64. Wei, C., Liu, Q., Ye, Z., Wang, Q., Wang, X., Wan, P., Gai, K., Chen, W.: Univideo: Unified understanding, generation, and editing for videos. arXiv preprint arXiv:2510.08377 (2025)
65. Wu, Y., Chen, L., Li, R., Wang, S., Xie, C., Zhang, L.: Insvie-1m: Effective instruction-based video editing with elaborate dataset construction. In: ICCV. pp. 16692–16701 (2025)
66. Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., Dubnov, S.: Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In: ICASSP (2023)
67. Xing, Y., He, Y., Tian, Z., Wang, X., Chen, Q.: Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In: CVPR. pp. 7151–7161 (2024)
68. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
69. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al.: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2024)
70. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: MANIQA: Multi-dimension attention network for no-reference image quality assessment. In: CVPR (2022)
71. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
72. Yuan, L., Wang, J., Sun, H., Zhang, Y., Lin, Y.: Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. arXiv preprint arXiv:2501.07888 (2025)
73. Zhang, G., Zhou, Z., Hu, T., Peng, Z., Zhang, Y., Chen, Y., Zhou, Y., Lu, Q., Wang, L.: UniAVGen: Unified audio and video generation with asymmetric cross-modal interactions. arXiv preprint arXiv:2511.03334 (2025)

74. Zhang, Y., Yang, H., Zhang, Y., Hu, Y., Zhu, F., Lin, C., Mei, X., Jiang, Y., Peng, B., Yuan, Z.: Waver: Wave your way to lifelike video generation. arXiv preprint arXiv:2508.15761 (2025)
75. Zhang, Y., Gu, Y., Zeng, Y., Xing, Z., Wang, Y., Wu, Z., Chen, K.: Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. arXiv preprint arXiv:2407.01494 (2024)
76. Zhao, L., Feng, L., Ge, D., Chen, R., Yi, F., Zhang, C., Zhang, X.L., Li, X.: UniForm: A unified multi-task diffusion transformer for audio-video generation. arXiv preprint arXiv:2502.03897 (2025)
77. Zheng, H., Weng, S., Liu, J., Yang, S., Shi, B., Wang, X.: Audio-sync video instance editing with granularity-aware mask refiner. arXiv preprint arXiv:2512.10571 (2025)